Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells

Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells

Zhixing Cao[a,b] and Ramon Grima[b,1]

[a]The Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237, People's Republic of China; and [b]School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3BF, United Kingdom

Edited by Charles S. Peskin, New York University, New York, NY, and approved January 21, 2020 (received for review June 25, 2019)

The stochasticity of gene expression presents significant challenges to the modeling of genetic networks. A two-state model describing promoter switching, transcription, and messenger RNA (mRNA) decay is the standard model of stochastic mRNA dynamics in eukaryotic cells. Here, we extend this model to include mRNA maturation, cell division, gene replication, dosage compensation, and growth-dependent transcription. We derive expressions for the time-dependent distributions of nascent mRNA and mature mRNA numbers, provided two assumptions hold: 1) nascent mRNA dynamics are much faster than those of mature mRNA; and 2) gene-inactivation events occur far more frequently than gene-activation events. We confirm that thousands of eukaryotic genes satisfy these assumptions by using data from yeast, mouse, and human cells. We use the expressions to perform a sensitivity analysis of the coefficient of variation of mRNA fluctuations averaged over the cell cycle, for a large number of genes in mouse embryonic stem cells, identifying degradation and gene-activation rates as the most sensitive parameters. Furthermore, it is shown that, despite the model's complexity, the time-dependent distributions predicted by our model are generally well approximated by the negative binomial distribution. Finally, we extend our model to include translation, protein decay, and auto-regulatory feedback, and derive expressions for the approximate time-dependent protein-number distributions, assuming slow protein decay. Our expressions enable us to study how complex biological processes contribute to the fluctuations of gene products in eukaryotic cells, as well as allowing a detailed quantitative comparison with experimental data via maximum-likelihood methods.

stochastic gene expression | master equation | perturbation theory

In the past two decades, advances in the real-time measurement of single-cell dynamics have revealed the stochastic nature of gene expression (1) and spurred a huge interest in the construction, simulation, and analytic solution of stochastic models of intracellular processes (2–4). Many experiments report the measurement of messenger RNA (mRNA), and, hence, there is a general need for stochastic models which can realistically predict the temporally varying distribution of mRNA molecule numbers in single cells. The word "realistically" is key because while there are a number of stochastic models of mRNA fluctuations in the literature, nevertheless, because of the complexity of the mRNA life cycle, currently very few of these models incorporate some of the detailed biological knowledge gleaned from single-cell experiments—this is particularly true for eukaryotic cells, where the transcription process is more complex than in prokaryotes and where compartmentation plays an important role (5).

The simplest stochastic model of mRNA fluctuations assumes that the gene is continuously ON, producing mRNA at some constant rate, followed by mRNA decay or its dilution due to cell division (often referred to as a constitutive expression model). If all these processes are approximated by effective first-order reactions, then the model is easy to solve and predicts a Poisson distribution of mRNA molecule numbers in cells (6). However, there is a large body of experimental evidence showing that the distribution of molecule-number fluctuations is typically non-Poisson (7–10) [except for housekeeping genes (11)], and, hence, modifications of this model are clearly needed. Adding an intermediate state which can either represent nascent mRNA or nuclear mRNA leads to the same Poisson distribution (12) (M1 in Fig. 1*A*). In contrast, assuming that a gene can switch between an ON and an OFF state (M2 in Fig. 1*A*) does lead to non-Poissonian mRNA fluctuations. The model has also been solved exactly analytically (13), and, in certain limits, it predicts bursty mRNA expression (8), a phenomenon which has been measured experimentally (14). This model, commonly called the two-state or telegraph model, has thus been widely adopted in the literature as the standard model of stochastic mRNA dynamics in eukaryotic cells (15–17). A recent application of the model is its use to infer the promoter-switching rates and the transcription rate of thousands of genes in mouse and human fibroblasts (10) from single-cell RNA-sequencing data. However, it is clear that this model is still far from including well-known processes, such as mRNA maturation, cell division, gene replication, dosage compensation, and growth-dependent transcription, all of which have been shown to have significant effects on the mRNA molecule numbers inside cells.

Significance

The random nature of gene expression is well established experimentally. Mathematical modeling provides a means of understanding the factors leading to the observed stochasticity. In this article, we extend the classical two-state model of stochastic mRNA dynamics to include a considerable number of salient features of single-cell biology, such as cell division, replication, mRNA maturation, dosage compensation, and growth-dependent transcription. By means of biologically relevant approximations, we obtain expressions for the time-dependent distributions of mRNA and protein numbers. These provide insight into how fluctuations are modified and controlled by complex intracellular processes.

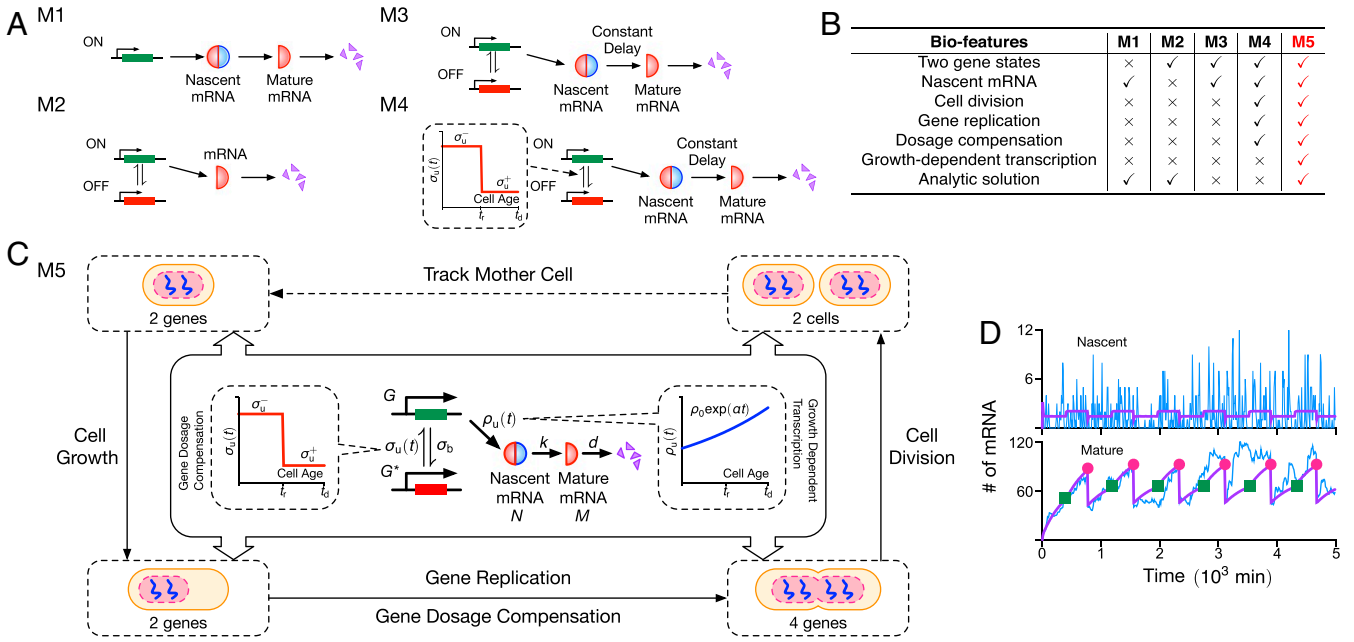
Author contributions: Z.C. and R.G. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: ramon.grima@ed.ac.uk.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1910888117/-/DCSupplemental.



Bio-features	M1	M2	M3	M4	M5
Two gene states	×	✓	✓	✓	✓
Nascent mRNA	✓	×	✓	✓	✓
Cell division	×	×	×	✓	✓
Gene replication	×	×	×	✓	✓
Dosage compensation	×	×	×	✓	✓
Growth-dependent transcription	×	×	×	×	✓
Analytic solution	✓	✓	×	×	✓

Fig. 1. (*A*) Illustration of four models of stochastic mRNA dynamics in the literature. Note that nascent mRNA is shown as joined blue and red semicircles, illustrating its unspliced nature (blue for introns and red for exons), while mature mRNA being composed of only exons is shown as red semicircles. The models describe transcription (constitutive, e.g., M1; or intermittent, e.g., M2, M3, and M4), mRNA maturation (M1, M3, and M4), and details of the cell cycle (M4); their biological features are compared in *B* and discussed in the main text. Only the simplest of these two models (M1 and M2) have been analytically solved. The model (M5) proposed in this article is illustrated in *C*: It builds upon model M4 by adding growth-dependent transcription, describes maturation as a stochastic process, and has the major advantage of being analytically solvable. The model is composed of nonreactive components (dosage compensation, replication, cell division, and growth-dependent transcription) and reactive components; the latter are shown in the central boxes where G, G^* denote genes in the ON, OFF states; N is nascent mRNA; and M is mature mRNA. (*D*) We show stochastic simulations of M5 using the SSA (25), where the purple lines denote the mean, and a typical time series is shown in blue. The green squares and red dots indicate the gene-replication time (t_r) and cell-division time (t_d), respectively. We use parameters measured for *Nanog* in mouse embryonic stem cells from ref. 26: $\rho_u = 2.11$ min^{-1}, $d = 0.00245$ min^{-1}, $\sigma_b = 0.609$ min^{-1}, and $\sigma_u^- = 0.0282$ min^{-1}. The gene-replication time $t_r = 400$ min, cell-division time $t_d = 780$ min, maturation rate $k = 0.1299$ min^{-1}, and gene-dosage parameters $\sigma_u^+ = 0.71\sigma_u^-$ are reported in ref. 19 for the same type of cells. Note that we set $\alpha = 0$, meaning that there is no growth-dependent transcription. Each realization is initiated with zero nascent and mature mRNAs and the gene in the ON state.

There are few studies which have modified the standard model of stochastic mRNA dynamics to include some of the mentioned biological processes: Senecal et al. (18) modified the standard model by including the conversion of nascent mRNA to mature mRNA after a constant time delay (model M3 in Fig. 1*A*), while Skinner et al. (19) extended this model further by including cell division, gene replication, and dosage compensation (model M4 in Fig. 1*A*). Note that there are also models which explicitly account for cell division (20, 21) or for replication (22), but have no description for promoter switching or other detailed features of the mRNA life cycle.

The disadvantage of the last two models (M3 and M4) compared to the previous two (M1 and M2) is that there is no known analytic solution to them. Hence, they have been explored purely via stochastic simulations. In Fig. 1*B*, we summarize the features of the models that we have discussed. In this paper, we develop an analytically tractable stochastic model of mRNA dynamics in diploid cells, which includes all of the processes in the most advanced simulation model (M4) developed thus far, while also additionally including growth-dependent transcription (the scaling of transcription with cellular volume) (23), a mechanism maintaining mRNA concentration homeostasis (24) irrespective of changes in cellular volume and DNA content. We also extend the model to include protein dynamics. The advantage of our theory is that it leads to intuitive insight that cannot be easily obtained from stochastic simulations. In addition, our model provides a framework which can be easily extended to include more intricate biological phenomena.

Results

Model Specification. The dynamic biochemical processes described by our model are illustrated in Fig. 1*C* and are as follows:

1) In the prereplication stage of the cell cycle, two gene copies can independently switch from ON (G) to OFF state (G^*) with rate σ_b and from OFF to ON with rate σ_u^-. Gene-copy independence has been shown experimentally (19).
2) In the ON state, there is the production of nascent (unspliced mRNA) denoted by N which subsequently becomes mature (spliced; denoted by M) with rate k. We assume that the maturation time is exponentially distributed rather than a deterministic time delay as assumed in previous studies (18, 19). This assumption is used because it makes the model analytically tractable (since then the model is Markov) and also since it has been shown (28) that the distribution of nascent mRNA numbers is insensitive to whether the delay is deterministic or exponentially distributed with identical mean maturation time, provided that $\rho_u \gg \sigma_b, \sigma_u$, which is the case experimentally (Table 1).
3) The mature mRNA decays with rate d via first-order kinetics, which is a common assumption supported by experiments (29).
4) Growth-dependent transcription whereby the transcription rate is proportional to the volume of the cell—this models the mRNA-concentration homeostasis mechanism reported in ref. 23. We assume that the cell volume increases exponentially with cell age t

Table 1. Kinetic parameters reported in various experimental papers

Cell type (gene)	σ_u(min^{-1})	σ_b(min^{-1})	ρ_u (min^{-1})	d (min^{-1})	Burst size (ρ_u/σ_b)	Fraction ON time ($\sigma_u/(\sigma_u+\sigma_b)$)	Timescale ratio δ	Reference
Yeast (POL1)	0.0700	0.680	2.00	0.0693	2.9	0.093	9.71	(11)
Yeast (PDR5)	0.3000	5.300	11.30	0.0495	2.1	0.054	17.67	(11)
Mouse embryonic stem cells (Oct4)	0.0092	0.018	1.90	0.0023	105.6	0.338	1.96	(19)
Mouse embryonic stem cells (Nanog)	0.0019	0.007	0.80	0.0022	115.9	0.216	3.63	(19)
Mouse embryonic stem cells (Nanog)	0.0282	0.609	2.11	0.0025	3.5	0.044	21.60	(26)
Human osteosarcoma (c-Fos)	0.1075	0.313	7.30	0.0462	23.4	0.256	2.91	(18)
Mouse hepatocytes (Acly)	0.0010	0.002	0.25	0.0004	129.3	0.337	1.97	(27)
Mouse hepatocytes (Actb)	0.0013	0.036	2.52	0.0004	70.1	0.034	28.77	(27)
Mouse hepatocytes (Srebf1)	0.0015	0.013	1.80	0.0022	137.7	0.102	8.82	(27)
Mouse hepatocytes (Insr1)	1.6×10^{-5}	3.5×10^{-4}	0.03	3.3×10^{-5}	81.0	0.045	21.00	(27)
Mouse fibroblasts (3,575 genes)	0.0022	0.236	0.69	0.0035	6.9	0.092	101.73	(10)
Human fibroblasts (1,609 genes)	0.2173 (d)	6.752 (d)	272.11 (d)	N/A	134.5	0.102	34.86	(10)
Mouse fibroblasts (16 genes)	0.0136	0.167	2.48	0.0109	17.2	0.074	20.11	(14)

The transcription rate (ρ_u), the mRNA degradation rate (d), the rate at which the gene switches from ON to OFF (σ_b), and the rate at which it switches from OFF to ON (σ_u) have been estimated from experimental data by using various models of stochastic mRNA dynamics (mostly using the standard model M2 in Fig. 1*A*). The data reveal that gene expression is bursty (gene is ON for short times, and, in that time, a large burst of mRNA produced). The large values of the ratio $\delta=\sigma_b/\sigma_u$ show that gene-inactivation events occur far more frequently than gene-activation events. Note that 1) estimates for the last three rows represent averages over genes; 2) human fibroblast estimates for σ_u, σ_b, ρ_u are in terms of d, and the latter was not measured; and 3) we do not report the value of σ_u before and after replication because the vast majority of studies do not take into account replication and, hence, report a single value. See *SI Appendix*, Section 2 for details. N/A, not applicable.

(where $0\le t\le t_d$ and t_d is the cellular interdivision time); this assumption is supported by experimental evidence for a variety of mammalian cells (30). Hence, the effective transcription rate follows the equation $\rho_u(t)=\rho_0\exp(\alpha t)$, where ρ_0 is the transcription rate at the start of the cell cycle, $\alpha=(1/t_d)\ln(V_f/V_0)$; V_0 is the cell volume at the beginning of the cell cycle; and V_f is the cell volume just before the cell divides. If there is no growth-dependent transcription, then $\rho_u(t)$ is a time-independent constant and corresponds to setting $\alpha=0$.

5) Replication results in a doubling of the gene copies from two to four at cell age t_r (replication time). We assume that this occurs instantaneously—i.e., replication occurs over a period which is much shorter than the length of the cell cycle. We shall refer to the gene which is replicated as the mother copy and the duplicated gene as the daughter copy. The daughter copy can either inherit the gene state from the mother copy (31), or else all copies (mother and daughter) are reset to the OFF state upon replication. One plausible explanation for the latter case is that to avoid the potential risk of conflict between replication and transcription (and the resulting DNA damage), it is highly likely that in the region where replication is actively ongoing or just completed, there is no transcription, indicating an OFF state (figure 2C in ref. 32).
6) Dosage compensation is modeled as a change in the value of the rate at which the gene switches from OFF to ON upon replication, specifically $\sigma_u=\sigma_u^-$ if $0\le t<t_r$ and $\sigma_u=\sigma_u^+$ if $t_r\le t\le t_d$. This assumption can explain experimental data (19). Note that dosage compensation is another mechanism (besides growth-dependent transcription) which results in approximate mRNA concentration homeostasis (24) over the duration of the cell cycle (33).
7) Binomial partitioning of nascent and mature mRNA at cell division. We here assume that nascent and mature mRNA segregate independently of each other with a probability $1/2$ of ending up in one of the two daughter cells. The time between successive cell divisions is assumed to be fixed. This is a simplification, since a number of experiments show interdivision time variability (34, 35). The assumption of a fixed cell-cycle length is made to make the mathematical analysis of the model tractable. We will also show later how the theory can be modified to describe the effect of cell-cycle-length variability.

Approximate Solution of the Model. A master equation can be written which describes the exact stochastic dynamics of the above model with the replication and cell-division processes modeled via appropriate boundary conditions (see *SI Appendix*, section 1 for details). Given the myriad complex biological functions described by the model, it should come as no surprise that we were unable to find an exact solution to this master equation (indeed, much simpler models often cannot be solved exactly; see ref. 4 for a review of the state of the art in solutions of chemical master equations). Our approach will consist of breaking the model into submodels, where each considers only a subset of bioprocesses, followed by solving each submodel approximately and then integrating the results to obtain a solution to the full model.

We note from Table 1 that the vast majority of eukaryotic genes are characterized by a large value of the ratio $\delta=\sigma_b/\sigma_u$ (gene-inactivation rate divided by the gene-activation rate), i.e., genes spend most of their time in the OFF state. For the moment, we ignore the processes of cell division and replication and focus on nascent mRNA dynamics due to promoter switching, growth-dependent transcription, and maturation for a single gene copy. As we show in *SI Appendix*, Section 3.1, in this case for large δ, the generating function corresponding to the time-dependent marginal distribution of nascent mRNA numbers of a single gene $P(n_N,t)$ can be written (by a slight abuse of notation) as:

$$G(u,t)=\sum_{n_N=0}^{\infty}(1+u)^{n_N}P(n_N,t) \qquad [1]$$
$$=g(ue^{-kt})\left(\frac{\rho_0 e^{-kt}u-\sigma_b}{\rho_0 e^{\alpha t}u-\sigma_b}\right)^{\frac{\sigma_u}{\alpha+k}}.$$

Here, we have assumed that the initial marginal distribution of i nascent mRNA molecules is $P(i)=p_i$, which implies $g(u)=\sum_i p_i(u+1)^i$. It can be shown that Eq. **1** implies that for large δ, the stochastic reaction dynamics stemming from the combined processes of promoter switching and nascent mRNA production with a time-dependent transcription rate to account for growth-dependent transcription can be described by a simpler system of one effective reaction. In this reaction, nascent mRNA is produced at rate σ_u in bursts whose size are distributed according to a negative binomial distribution with a time-dependent mean burst size $(\rho_0/\sigma_b)\exp(\alpha t)$ (*SI Appendix*, Section 3.2). The major advantage of this effective reaction description is that it dispenses with an explicit gene-state description which considerably simplifies the calculations to follow. In particular, the issue of how to choose the gene state at the beginning of replication is circumvented—intuitively, this is possible because since the gene spends most of its time in the OFF state, the two mechanisms described in point 5 in Model Specification cannot be distinguished in practice.

$$G_A(u,t)=\begin{cases}\left(\dfrac{\rho_0 u e^{-kt}-\sigma_b}{\rho_0 u e^{\alpha t}-\sigma_b}\right)^{\frac{\sigma_u^-}{\alpha+k}} & t\in[0,t_r),\\ \left(\dfrac{\rho_0 u e^{-kt}-\sigma_b}{\rho_0 u e^{-k(t-t_r)+\alpha t_r}-\sigma_b}\right)^{\frac{\sigma_u^-}{\alpha+k}}\left(\dfrac{\rho_1 u e^{-k(t-t_r)}-\sigma_b}{\rho_1 u e^{\alpha(t-t_r)}-\sigma_b}\right)^{\frac{\sigma_u^+}{\alpha+k}} & t\in[t_r,t_d),\end{cases} \quad \textbf{[2A]}$$

$$G_B(u,t)=\begin{cases}1 & t\in[0,t_r),\\ \left(\dfrac{\rho_1 u e^{-k(t-t_r)}-\sigma_b}{\rho_1 u e^{\alpha(t-t_r)}-\sigma_b}\right)^{\frac{\sigma_u^+}{\alpha+k}} & t\in[t_r,t_d).\end{cases} \quad \textbf{[2B]}$$

Next, we include the processes of replication and dosage compensation. Specifically, let $G_A(u,t)$ and $G_B(u,t)$ be the generating functions describing the dynamics of nascent mRNAs born within a cell cycle for a single mother and daughter copy, respectively. By using Eq. **1**, it follows that the generating functions of nascent mRNA produced by mother and daughter copies are piece-wisely defined and given by Eqs. **2A** and **2B**. The first part $t\in[0,t_r)$ of $G_A(u,t)$ describes the stochastic dynamics of nascent mRNA born in the prereplication time. Note that the initial condition is zero, and $\sigma_u=\sigma_u^-$. The second part $t\in[t_r,t_d)$ of $G_A(u,t)$ describes the stochastic dynamics of nascent mRNA born in the postreplication time. This is given by Eq. **1**, with g replaced by the initial condition which is the generating function at replication time (from the expression for $t\in[0,t_r)$); also note that $\sigma_u=\sigma_u^+$ (due to dosage compensation), ρ_0 is replaced by $\rho_1=\rho_0 e^{\alpha t_r}$ since this is the transcription rate at replication time, and time t is replaced by $t-t_r$. Note that, intuitively, the second part $t\in[t_r,t_d)$ of $G_A(u,t)$ (which describes postreplication dynamics) can be written as a product of two factors because there is independence between nascent mRNA inherited from the prereplication stage and the nascent mRNA born in the postreplication stage. Since there is no transcription activity in the prereplication time for the daughter copy, the generating function $G_B(u,t)$ is trivially equal to 1 for $t\in[0,t_r)$. The second part $t\in[t_r,t_d)$ of $G_B(u,t)$ can be found similarly as for $G_A(u,t)$. Note that the individual factors in the generating-function expressions can be written as a product of the generating functions for the binomial and negative binomial distributions (*SI Appendix*, section 3).

Finally, we add the details of cell division and the associated binomial partitioning. There are two processes contributing to the number of mRNAs at a particular cell age t of a given cell cycle n: 1) the decay of mRNAs inherited from the previous cycle, and 2) the production of new mRNAs in cell cycle n. These processes are independent from each other when the gene spends most of its time in one state (as in our case), and, hence, it follows that we can write:

$$\begin{aligned}P_n(n_N,t)&=\frac{1}{n_N!}\frac{d^{n_N}G_n(u,t)}{du^{n_N}}\bigg|_{u=-1},\\ G_n(u,t)&=\underbrace{G_n(ue^{-kt},0)}_{\text{death process}}\underbrace{G_A^2(u,t)G_B^2(u,t)}_{\text{new born mRNA}}\ \forall t\in[0,t_d),\end{aligned} \quad \textbf{[3]}$$

where $P_n(n_N,t)$ is the marginal distribution of nascent mRNA numbers at time t in cell cycle n. Note that the power of 2 on the right-hand side of Eq. **3** arises from the diploidy of gene copies and from assuming that they are independent of each other (as mentioned in point 1 of Model Specification). Binomial partitioning at cell age t_d leads to a relationship between the initial conditions for the generating function of the n^{th} cycle and the generating function of the $n-1^{\text{th}}$ cycle at time t_d, which can be shown (*SI Appendix*, section 3.3) to lead to the condition:

$$G_n(u,0)=G_{n-1}(\eta u,0)\,G_A^2\left(\frac{u}{2},t_d\right)G_B^2\left(\frac{u}{2},t_d\right),\ n>1 \quad \textbf{[4]}$$

where $\eta=e^{-kt_d}/2$. Note that $G_1(u,0)$ is the initial condition at the beginning of the first cycle, and this is a user-input condition; for all results in this paper, we assumed $G_1(u,0)=1$, implying no nascent mRNA initially.

Hence, summarizing Eqs. **3** and **4** together with Eqs. **2A** and **2B** provides an approximate time-dependent solution of the marginal distribution of nascent mRNA numbers valid for all cell ages and cellular generations; the assumption behind our derivation was that the gene spends most of its time in the OFF state. Note that while the derivation assumed a growth-dependent transcription rate as described in point 4 of Model Specification, nevertheless, it is straightforward to derive similar results for a general time-dependent transcription rate (*SI Appendix*, section 10); this may be useful to describe synthetic gene-regulatory networks where the transcription rate can be arbitrarily regulated over time.

In *SI Appendix*, section 4, we show that an approximate time-dependent solution of the marginal distribution of mature mRNA numbers can be similarly derived, provided that one further assumes that the timescales of nascent mRNA are much shorter than those of mature mRNA. The timescales of these two types of mRNA are determined by k and d, respectively (their elimination

rates), and there is strong evidence showing $k \gg d$. For the gene c-Fos in human osteosarcoma cells, Senecal et al. (18) estimated a value of $k = 1.25\ \text{min}^{-1}$, $d = 0.0462\ \text{min}^{-1}$, while for Nanog and Oct4 in mouse embryonic stem cells, Skinner et al. (19) estimated $k = 0.13\ \text{min}^{-1}$, $d = 0.0022\ \text{min}^{-1}$, and $k = 0.29\ \text{min}^{-1}$, $d = 0.0023\ \text{min}^{-1}$, respectively. Hence, k is considerably larger than d generally.

Specifically, we show using singular perturbation theory that if the gene spends most of its time in the OFF state ($\delta \gg 1$) and if nascent mRNA maturation is fast ($k \gg d$), then the marginal distribution of mature mRNA numbers within a single cell is approximately given by Eqs. **3** and **4** together with Eqs. **2A** and **2B**, with k replaced by d and with the change of variable $n_N \to n_M$. Note that under the fast-maturation assumption, the dynamics of nascent mRNA do not affect the dynamics of mature mRNA; for a detailed discussion, see *SI Appendix*, section 4.

Numerical Evaluation of the Accuracy of the Approximate Distributions. A main assumption behind our derivation is that $\delta \gg 1$ holds for many genes, but clearly this is not the case for all (Table 1). It is unclear how large δ has to be for our approximation to be accurate. Hence, we next test the accuracy of our theory by fixing $k \gg d$ (this assumption holds for all genes that we could find data for; *SI Appendix*, section 2) and varying parameters ρ_u and σ_b (relative to the rest which are fixed) to vary δ at constant mean burst size ρ_u/σ_b for the case of no growth-dependent transcription ($\alpha = 0$). We then quantify the accuracy of our approximation by calculating the Hellinger distance (HD) between the approximate probability distribution of nascent and mature mRNA numbers and the exact numerical solution of the chemical master equation as a function of δ, which is varied over two orders of magnitude from 2 to 200. Note that the HD has the properties of being symmetric and satisfies the triangle inequality, thus implying that it is a distance metric on the space of probability distributions (unlike, e.g., the commonly used Kullback–Leibler divergence); it is also conveniently a fraction, which makes for easy interpretation.

Our approximate theory was derived by using the assumption that $\delta \gg 1$, and, hence, we expect the accuracy of the theory to increase with δ. This is verified in the heatmap shown in Fig. 2*A*, where it is shown that the error in our approximate theory is inversely proportional to δ; is a weak function of absolute time, i.e., independent of cell age and generation; and is generally small (HD $\ll 1$) for both nascent (Fig. 2 *A*, *Upper*) and mature (Fig. 2 *A*, *Lower*) mRNA. The exact probability distributions for two time points are compared with the approximate distributions in Fig. 2*B* for three different values of δ. Note that the match between approximate analytic solution and the exact solution (using stochastic simulation algorithm [SSA]) is excellent for $\delta = 20$ and acceptable for $\delta = 5$ for all times; for smaller values of δ, the nascent mRNA distribution still is well approximated, but the same cannot be said for the mature mRNA distribution. Hence, our theory is accurate for the majority of genes reported in Table 1.

In Fig. 2*C*, we compare the time-dependent mean and variance predicted from the approximate theory with the exact result (from SSA), showing the accuracy of the theory to capture the cyclic behavior of the moments due to the dynamic processes of replication and cell division. The theory's accuracy remains high, even when growth-dependent transcription is turned on $\alpha > 0$ (*SI Appendix*, Fig. S2).

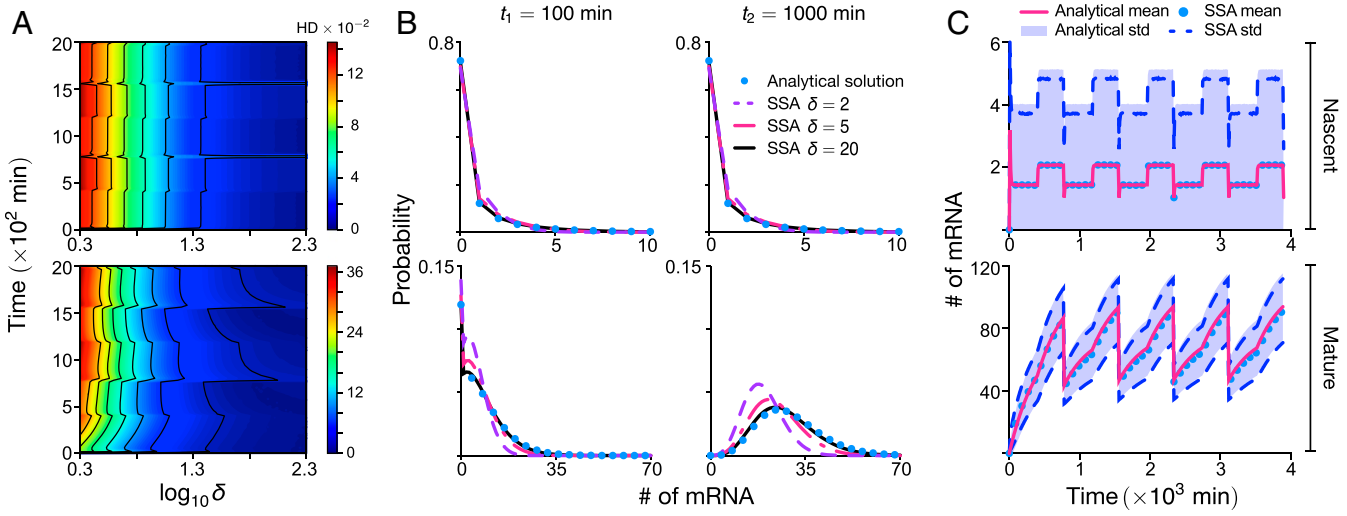


Fig. 2. Accuracy of the approximate analytic solution to the stochastic model of eukaryotic gene expression. The approximate probability-distribution solution for nascent mRNA numbers is given by Eqs. **3** and **4** together with Eqs. **2A** and **2B**; the distribution for mature mRNA numbers is given by the same equations with k replaced by d and n_N replaced by n_M. Here, we investigate the accuracy of this solution relative to the exact solution, which is numerically computed by using the Finite State Projection algorithm (FSP) or the SSA; note that these simulations are of the full model (without any approximation) as described in Model Specification and *SI Appendix*, section 1. *Upper* and *Lower* show information for nascent and mature mRNA, respectively. (*A*) The heatmap for the HD (which quantifies the distance between the approximate and exact marginal distributions) for both nascent and mature mRNA as a function of absolute time and the dimensionless timescale ratio δ. Note that $\delta = \sigma_b/\sigma_u^-$ is here varied by changing σ_b while keeping $\sigma_u^- = 0.0282\ \text{min}^{-1}$ and the mean burst size $\rho_u/\sigma_b = 3.5$ constant; the rest of the parameter values are the same as those in Fig. 1. Due to the computational demand of producing a heatplot, the exact solution is here computed by using FSP; computations using the SSA at random points in the heatplot were indistinguishable from those using FSP. (*B*) Marginal distributions of the nascent and mature mRNA for two time points and three values of δ compared to those obtained from stochastic simulations using the SSA. (*C*) Plots of the mean and SD (std) versus time as predicted by the approximate analytical solution and exact stochastic simulations for $\delta = 21.60$ using the SSA (which also agree with those using FSP). All plots show that the accuracy of the distributions and moments predicted by the approximate analytic solution is high provided δ is not too small (larger than about five). Note that $\alpha = 0$ in all panels, meaning that there is no growth-dependent transcription.

Effect of Cell-Cycle-Length Variability on mRNA Distributions. Our theory assumes a fixed cell-cycle length and synchronized cell cycles among cells. This is the case when cells are subjected to certain environmental conditions (36, 37), when the circadian clock gates the cell cycle (38, 39), and during certain phases of morphogenesis (40). However, variation in the cell-cycle length is likely common (e.g., in Fig. 3*A*, we show experimental data for mouse fibroblasts), which leads to asynchronous behavior. We modified the SSA such that the cell-cycle times t_d are assumed to be random variables independently drawn from an Erlang distribution (which well approximates the experimental data in Fig. 3*A*) and the replication time is exactly in the middle of each cycle. Each trajectory of the algorithm corresponds to a forward lineage, i.e., either of the two daughter cells is followed with equal probability. The distribution of the number of mature mRNAs is constructed from an ensemble of these single-cell trajectories; this distribution is shown as blue dots in Fig. 3*B*, where the parameters are those measured for two mouse genes. The mature mRNA distribution can also be predicted by modifying our theory to take into account asynchronous cell cycles, but keeping the assumption of fixed cell-cycle length (*SI Appendix*, section 6); this prediction is shown as a red curve in Fig. 3*B*. Excellent agreement between theory and the SSA is found for 16 mouse genes (two are shown in Fig. 3*B* and the rest in *SI Appendix*, Fig. S5). As shown in *SI Appendix*, section 6, the implicit reason for the accuracy of the modified theory is the fact that the mRNA lifetime in mouse fibroblasts is typically much less than the average cell-cycle length; i.e., rapid degradation averages out timing fluctuations, which is in line with other recent studies (41).

Sensitivity Analysis of the Coefficient of Variation of mRNA Fluctuations. An important use of the analytic results is that we can efficiently calculate the sensitivity of the coefficient of variation of mature mRNA (SD divided by the mean) to small perturbations in the eight parameters of the model. To this end, we first used our approximate theory to calculate closed-form expressions for the cyclo-stationary mean and variance of mature mRNA, which we denote as $\overline{\langle n_M \rangle}_t$ and $\overline{\sigma}^2_{n_M,t}$, respectively (*SI Appendix*, section 5). Note that the cyclo-stationary conditions ensue in the limit of biological steady-state growth (20), which is achieved when the probability that a cell of age t has a given number of molecules of certain species is independent of which generation it belongs to—i.e., setting $G_{n+1}(u,t) = G_n(u,t)$.

The cyclo-stationary coefficient of variation of mature mRNA noise averaged over the cell cycle is then given by $\overline{\mathrm{CV}} = t_d^{-1} \int_{t=0}^{t_d} \sqrt{\overline{\sigma}^2_{n_M,t}} / \overline{\langle n_M \rangle}_t \, dt$ (which is computed numerically over 100 discrete time points evenly distributed within a cell cycle). The relative sensitivity of $\overline{\mathrm{CV}}$ to a parameter r is then given by $\Lambda_r = (r/\overline{\mathrm{CV}}) \partial \overline{\mathrm{CV}} / \partial r$ (43), meaning that a 1% change in the value of parameter r leads to Λ_r% change in $\overline{\mathrm{CV}}$. We next computed the relative sensitivities for eight of the rate parameters for a large

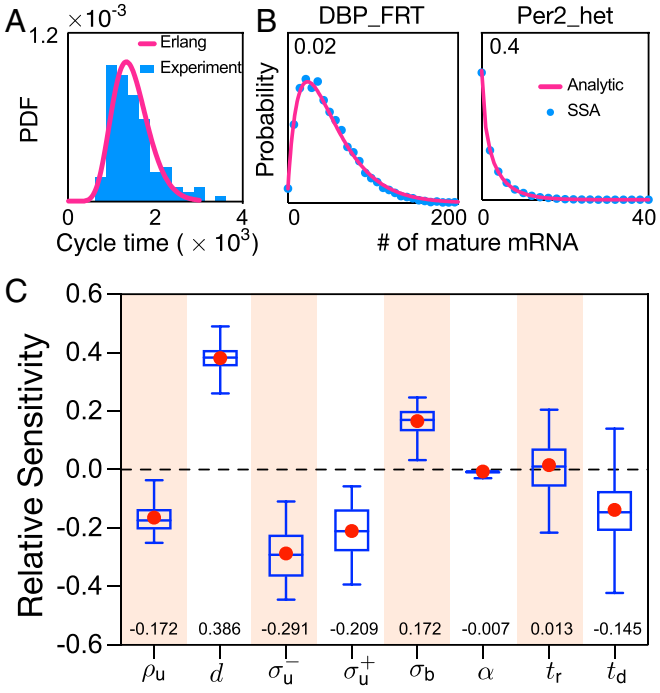


Fig. 3. (*A* and *B*) Cell-cycle-length variability and its effect on mature mRNA distributions. (*A*) The Erlang distribution provides a good fit to the experimentally measured cell-cycle time distribution of NIH 3T3 cells in ref. 42. PDF, probability distribution function. (*B*) The SSA modified such that the cell-cycle times are random variables independently drawn from an Erlang distribution is used to obtain the mature mRNA distributions (for the genes DBP_FRT and Per2_het reported in *SI Appendix*, Table S1) measured across an ensemble of cells (blue dots; see main text for details). The mature mRNA distributions are accurately predicted by modifying our theory (red solid lines; *SI Appendix*, section 6, Eq. 26) to take into account the asynchronicity of cell cycles across the population. Note that replication always occurs in the middle of a cell cycle. (*C*) Relative sensitivity of the cyclo-stationary coefficient of variation of mature mRNA noise averaged over the cell cycle to eight parameters. Box plots indicate the median and the 25% and 75% quantiles, with the mean marked as red dots. The median relative sensitivities are also shown as numbers at the bottom of the plot. The sensitivity analysis is carried out on 567,540 parametric combinations estimated for 1,051 genes of CAST allele data of mouse embryonic stem cells. Our results show that the degradation rate d is the most sensitive parameter, followed by σ_u^-, while α is the least sensitive one. Note that there is no dependence of the coefficient of variation of mature mRNA on k in the approximate theory, and, hence, k is not a relevant parameter. See *SI Appendix*, section 11 for the specific range of parameter choice and their justification.

number of genes in mouse embryonic stem cells (10). Note that the 1,051 genes selected for this analysis are characterized by a timescale ratio $\delta \geq 5$, a value which is large enough to guarantee the accuracy of our approximate analytic solution (Fig. 2), which is used to calculate the coefficient of variation. The results illustrated by using box plots in Fig. 3*C* show that the most sensitive parameters were the mRNA degradation rate d and the dosage-compensation parameters σ_u^-, σ_u^+, while the least sensitive parameters were the growth-dependent transcription parameter α and the replication time t_r.

Effective Negative Binomial Approximation of Mature mRNA Distributions. While our theory gives approximate distributions of nascent and mRNA molecule numbers, these distributions are complex and cannot be easily written in terms of known simple distributions. It has been frequently observed that many measured number distributions can be easily fit by the negative binomial distribution (or its continuous analog the gamma distribution). Indeed, this is one of the major reasons why the two-state model of mRNA dynamics (model M2 in Fig. 1) has become widely adopted, since in the bursty limit, the probability distribution of molecule numbers is negative binomial. Hence, we next investigate whether our approximate distributions can also be well fit by negative binomial distributions.

Assuming a negative binomial distribution (NB) for the number of molecules of mature mRNA, $P(n_M) \sim \mathrm{NB}(r, p)$ with parameters r and p, its mean and variance are given by $\langle n_M \rangle_e = rp/(1-p), \sigma^2_{n_M,e} = rp/(1-p)^2$. We equate these two moments to the cyclo-stationary moments from our theory: Specifically, we set $\langle n_M \rangle_e = \overline{\langle n_M \rangle}_t/2$ and $\sigma^2_{n_M,e} = \overline{\sigma}^2_{n_M,t}/2$, where the factor of $1/2$ accounts for the two independent gene copies in our model. One can then find simple expressions for r and p:

$$r = \frac{1}{2}\frac{\overline{\langle n_M \rangle}_t^2}{\overline{\sigma}^2_{n_M,t} - \overline{\langle n_M \rangle}_t}, \; p = 1 - \frac{\overline{\langle n_M \rangle}_t}{\overline{\sigma}^2_{n_M,t}}.$$

Hence, we have constructed a negative binomial approximation $\mathrm{NB}(r, p)$ to our model's predicted distribution of the mature mRNA number distribution at cell age t in the cyclo-stationary limit. We test the accuracy of this approximation in Fig. 4. In Fig. 4*A*, we compute the HD distance between the distribution solution of our model (as computed by using Eqs. **2A**, **2B**, **3**, and **4** with k changed to d and n_N changed to n_M) and the negative binomial approximation for 21 time points in the cell cycle using parameters for 1,051 genes in mouse embryonic stem cells (same as used for sensitivity analysis). Remarkably, we find the HD to be much <1 for all genes and all cell ages, implying that the negative binomial approximation is an excellent one in practice. Given the well-known fact that the negative binomial is a good approximation to the steady-state solution of the standard two-state model of mRNA dynamics in bursty conditions (8), in an indirect sense, our results in Fig. 4*A* also show that the cyclo-stationary distribution of our model at a particular cell age can be well approximated by the steady-state distribution of the two-state model (for a particular choice of effective parameters). This argument is further reinforced in Fig. 4*B*, where we show that the HD averaged over one cycle for a particular gene is roughly linearly dependent with Θ in log space, where Θ is the absolute difference between the uncentered third moments of the two-state model and its negative binomial approximation. Note that Θ is defined in terms of the parameters ρ_u, σ_u^-, and d specific to a particular gene:

$$\Theta = \frac{2\hat{\rho}_u^3 \hat{\sigma}_b \hat{\sigma}_u^- (1+\hat{\sigma}_u^-)}{(\hat{\sigma}_b + \hat{\sigma}_u^-)^2 (1+\hat{\sigma}_b+\hat{\sigma}_u^-)^2 (2+\hat{\sigma}_b+\hat{\sigma}_u^-)}, \qquad [5]$$

where $\hat{\rho}_u = \rho_u/\hat{d}$, $\hat{\sigma}_u^- = \sigma_u^-/\hat{d}$, $\hat{\sigma}_b = \sigma_b/\hat{d}$, and $\hat{d} = d + \ln 2/t_d$. For a derivation of this expression, see *SI Appendix*, section 7. Eq. **5** shows that the error in the negative binomial approximation is inversely proportional to the rate of promoter switching and directly proportional to the transcription rate. In Fig. 4*C*, we show the full versus effective negative binomial distributions as a function of cell age for the gene with the largest HD (Ndufa4)—even in this extreme case, the two distributions cannot be distinguished by eye, thus showing the high accuracy of the negative binomial approximation to our model for a large number of genes in mouse embryonic stem cells.

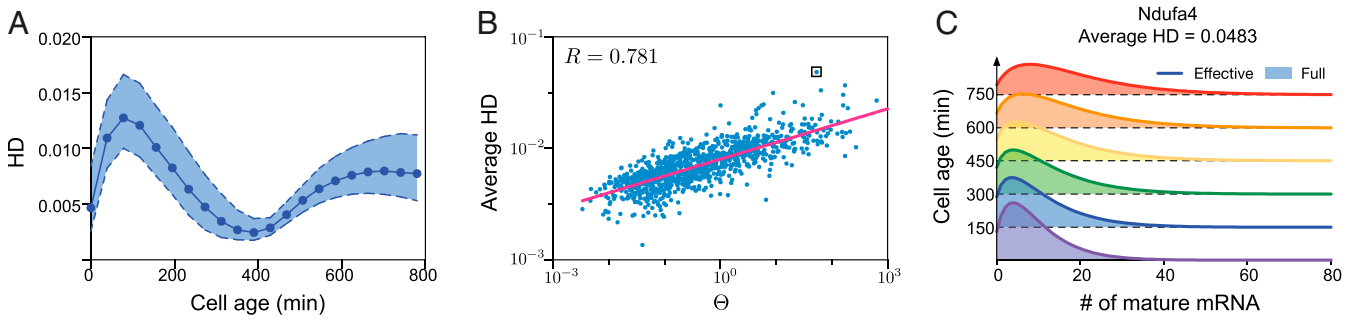


Fig. 4. Effective negative binomial approximation for the mature mRNA number distribution of our model. The approximation is obtained by matching the cyclo-stationary mean and variance of our model to those of an effective negative binomial distribution. (*A*) We calculate the HD between the effective distribution and the full distribution for 21 equidistant time points through a whole cell cycle. Each point corresponds to the HD median for 1,051 genes in mouse embryonic stem cells (same data used for Fig. 3*C*), whereas the broken lines show the 25% and 75% quantiles. Note that the four parameters t_d = 780 min, $\alpha = 0$, t_r = 400 min, and $\sigma_u^+ = 0.7\sigma_u^-$ are the same for all genes. (*B*) A plot of the HD for each gene averaged over 21 time points in the cell cycle versus the index Θ; the two quantities are linearly correlated in log space with Pearson correlation coefficient $R = 0.781$. (*C*) The matching of the effective and full distributions in time for the gene Ndufa4, which has the largest HD in *B* (shown as a boxed point).

Including Protein Dynamics. Thus far, the model only describes the stochastic mRNA dynamics. Given the increasing number of single-cell measurements of protein expression in eukaryotic cells (44–46), we next extend our theory to provide expressions for the protein distributions.

$$G_{\mathrm{A}}(u,t)=\begin{cases}\left[\dfrac{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u\mathrm{e}^{-d_{\mathrm{p}}t}-\sigma_{\mathrm{b}}}{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u-\sigma_{\mathrm{b}}}\right]^{\frac{\rho_{\mathrm{u}}\sigma_{\mathrm{u}}^{-}}{d_{\mathrm{p}}(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})}}, & t\in[0,t_{\mathrm{r}}),\\ \left[\dfrac{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u\mathrm{e}^{-d_{\mathrm{p}}t}-\sigma_{\mathrm{b}}}{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u\mathrm{e}^{-d_{\mathrm{p}}(t-t_{\mathrm{r}})}-\sigma_{\mathrm{b}}}\right]^{\frac{\rho_{\mathrm{u}}\sigma_{\mathrm{u}}^{-}}{d_{\mathrm{p}}(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})}}\left[\dfrac{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u\mathrm{e}^{-d_{\mathrm{p}}(t-t_{\mathrm{r}})}-\sigma_{\mathrm{b}}}{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u-\sigma_{\mathrm{b}}}\right]^{\frac{\rho_{\mathrm{u}}\sigma_{\mathrm{u}}^{+}}{d_{\mathrm{p}}(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})}} & t\in[t_{\mathrm{r}},t_{\mathrm{d}}),\end{cases} \quad \textbf{[6A]}$$

$$G_{\mathrm{B}}(u,t)=\begin{cases}1 & t\in[0,t_{\mathrm{r}}),\\ \left[\dfrac{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u\mathrm{e}^{-d_{\mathrm{p}}(t-t_{\mathrm{r}})}-\sigma_{\mathrm{b}}}{b(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})u-\sigma_{\mathrm{b}}}\right]^{\frac{\rho_{\mathrm{u}}\sigma_{\mathrm{u}}^{+}}{d_{\mathrm{p}}(\rho_{\mathrm{u}}+\sigma_{\mathrm{b}})}} & t\in[t_{\mathrm{r}},t_{\mathrm{d}}).\end{cases} \quad \textbf{[6B]}$$

The model has the same seven features as described at the beginning of Results, but with an additional two features: 1) mRNA is translated into protein at rate λ; and 2) protein decay occurs with rate d_{p}. Both reactions are assumed to obey first-order kinetics (6). Again, we need to make some approximation to proceed further: 1) We assume that the timescale ratio δ is large; and 2) we assume that the timescales of nascent and mature mRNA dynamics are much shorter than those of protein. The first assumption we know is satisfied for a large number of genes. The second assumption can be justified as follows. The timescales of nascent mRNA, mature mRNA, and protein are approximately given by the inverse of the elimination rates of each, i.e., k, d, and d_{p}, respectively. Now, as we saw earlier, $k \gg d$. Also, Schwanhäusser et al. (47) report that the median mRNA decay rate d is approximately five times larger than the median protein decay rate d_{p} for NIH 3T3 mouse fibroblasts (calculated over 4,200 genes); the cumulative distribution of the ratio of the two decay rates is shown in *SI Appendix*, Fig. S7. Similarly, for 1,962 genes in budding yeast, the median of the ratio of the mRNA decay rate to protein decay rate is approximately three (3). Hence, for a substantial number of genes, the assumption $k \gg d \gg d_{\mathrm{p}}$ holds, and that implies that protein dynamics occurs over a much slower timescale than both nascent and mature mRNA dynamics. Given assumptions 1 and 2, we can show using perturbation theory applied to the master equation of the model (*SI Appendix*, section 8) that the temporal protein distribution is given by Eqs. **3** and **4** with the replacements $k \to d_{\mathrm{p}}$ and $n_{\mathrm{N}} \to n_{\mathrm{P}}$ together with the generating functions given by Eqs. **6A** and **6B**, where $b = \lambda/d$ is the translational burst size quantifying the mean number of protein produced during the lifetime of mature mRNA. Note that this derivation is for the case of no growth-dependent transcription, i.e., $\alpha = 0$.

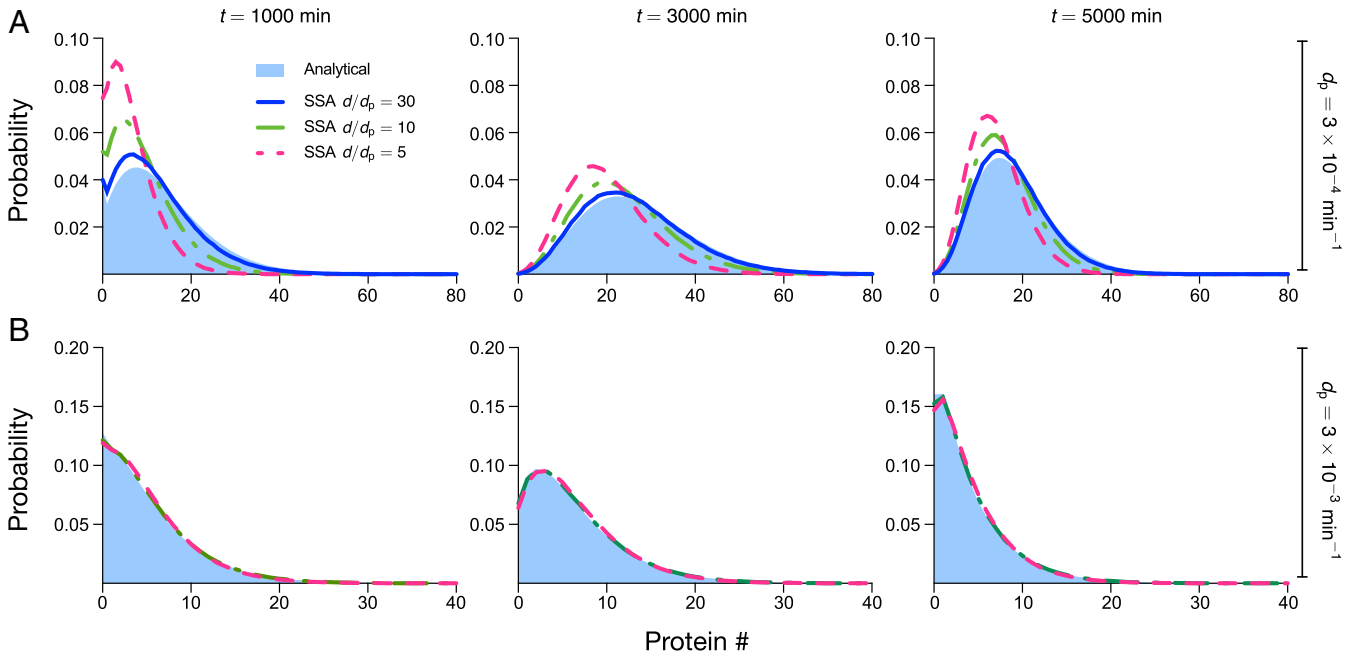


Fig. 5. Comparison of protein distributions predicted by theory (under the assumption of slow protein dynamics) and stochastic simulations using the SSA. (*A*) Using parameters typical of mammalian cells, we find that our theory agrees well with simulations for $d/d_{\mathrm{p}} = 30$, is acceptable for $d/d_{\mathrm{p}} = 10$, and performs poorly for $d/d_{\mathrm{p}} = 5$; agreement tends to be better with increasing time, but accuracy is mostly determined by d/d_{p}. Note that the protein lifetime is $\ln 2/d_p$ = 2,310 min, the cell-cycle duration is t_d = 1,560 min, and the mRNA lifetime is $\ln 2/d = 462$, 231, and 77 min for $d/d_{\mathrm{p}} = 5$, 10, 30, respectively. Given the value of t_{d}, we have that t = 1,000; 3,000; 5,000 min corresponds to cells in generations one, two, and three, respectively. (*B*) Parameter values as in *A*, except that the protein, mRNA decay, and translation rates are multiplied by 10. Note that the ratio d/d_{p} is unchanged from *A*, but both protein and mRNA lifetimes are now 10 times smaller, meaning that they are significantly less than the cell-cycle time. This condition leads to significantly improved agreement between theory and simulations, such that they are indistinguishable for $d/d_{\mathrm{p}} = 5$, 10, 30. See *SI Appendix*, section 12 for the choice of parameters.

The accuracy of the approximation is tested via stochastic simulations in Fig. 5*A* using parameters typical of mammalian cells (47). The protein distribution obtained from theory (shown as solid blue) is compared with SSA results for the protein distribution at three different times and three different values of the ratio of mRNA to protein-decay rates d/d_p. The discrepancy between theory and simulations decreases as d/d_p increases (as expected) and is particularly good for $d/d_p > 10$. There is also a small increase in accuracy of the theory with increasing absolute time, though the major determinant is the decay ratio. Simulations show that the accuracy of the theory increases if we increase the protein and mRNA decay rates while keeping their ratio constant—in particular, whenever the mRNA lifetime is much less than the protein lifetime and when the latter is less than the cell-cycle length, then the agreement with theory is excellent for a wide range of values of the ratio d/d_p (compare Fig. 5 *A* and *B*). Similarly, the accuracy of the theory also increases if we increase the cell-cycle length at constant ratio d/d_p [note that the limit of infinite cell-cycle length corresponds to the conventional case where partitioning due to cell division is not explicitly taken into account (3)]. Hence, summarizing, our expressions for the approximate protein distributions are accurate whenever $d/d_p \gg 1$ and $d_p t_d > 1$. From *SI Appendix*, Fig. S7, it can be deduced that about 20% of proteins in mammalian cells satisfy $d/d_p > 10$; also, analysis of the dataset in ref. 47 shows that only 30% of all proteins in mammalian cells have decay lifetimes less than the cell-cycle length. Hence, while timescale separation between mRNA and protein has played an important role in the development of reduced stochastic models of gene expression in bacteria and yeast (3), it appears that the same technique should be used with care when developing reduced models of stochastic gene expression in mammalian cells.

The model can also be further extended to include bimolecular gene–protein interactions, which are common in nature (48). Specifically, we consider the case of negative feedback mediated by an auto-regulatory motif, whereby the transition from the ON to OFF state of all gene copies (i.e., the two copies prereplication and the four copies postreplication) is mediated by protein binding to the gene. Using a recently developed technique, the linear mapping approximation (LMA) (49), we show in *SI Appendix*, section 9 how the distributions of mRNA and protein for the model with no feedback (derived earlier) can be used to construct approximate distributions for the model with feedback.

The LMA is based on two assumptions: 1) a conditional mean-field approximation which equates to assuming small protein fluctuations compared to the mean number of proteins when the promoter is unbound; and 2) a time-averaging assumption which corresponds to the first term of the Magnus expansion of the time-dependent solution of the master equation and which is uniformly valid in time, provided the protein–gene binding rates are not too large. The approximation error tends to be dominated by assumption 2. This, however, is typically small, as we show in Fig. 6, where it is clear that the LMA accurately captures the effect of negative feedback on the mature mRNA and protein distributions for both prereplication and postreplication times in the cell cycle.

Discussion

In this article, we have developed a model of gene expression in eukaryotic cells which includes a high level of biological detail compared to previous models in the literature, while remaining analytically tractable. Specifically, our model takes into account gene

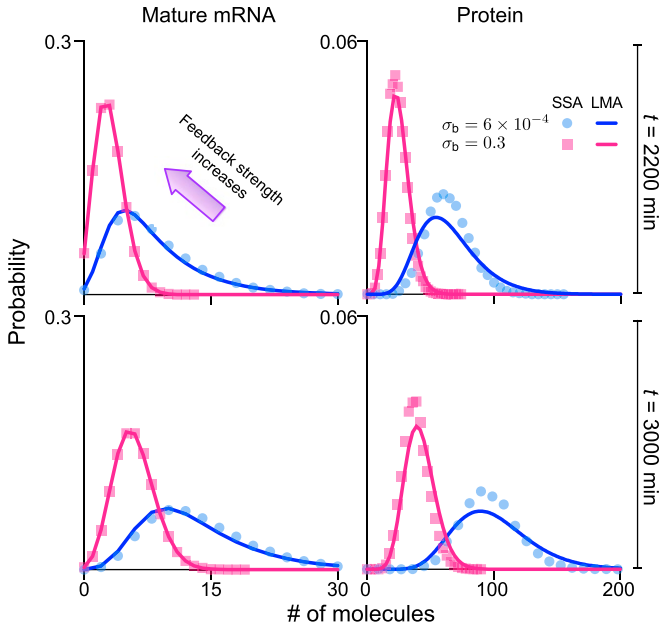


Fig. 6. LMA solution of the stochastic model with a negative-feedback auto-regulatory mechanism. The predicted distributions for mature mRNAs and proteins for time t = 2,200 min (prereplication) and t = 3,000 min (postreplication) of generation 2 are calculated from the equations for the generating functions in *SI Appendix*, section 9. LMA predictions for mature mRNA distributions agree with SSA results with remarkable accuracy, whereas the predictions for protein distributions agree with the SSA to an acceptable accuracy. Note that the ratio of decay rates d/d_p is five (the median reported in *SI Appendix*, Fig. S7), so that the timescale separation assumption is marginally satisfied. The graphs show that increasing feedback strength σ_b from 6×10^{-4} to 0.3 (500-fold change) substantially reduces the number of mature mRNAs and proteins. Specifically, the cell-cycle duration t_d = 1,560 min is selected to be 26 h, close to the data reported for NIH 3T3 in the supplementary information of ref. 47, the gene replication $t_r = 800$ min occurs roughly in the middle of the cell cycle. The decay rates of mature mRNA and protein are 0.005 and 0.001 min^{-1}, respectively (the half-lives are 2.3 and 11.5 h, respectively) and, hence, within the range for the same cell line reported in figure 2C of ref. 47. The other kinetic parameters are chosen as $\rho_u = 0.15\ \text{min}^{-1}$, $\rho_b = 0.0075\ \text{min}^{-1}$, $k = 10\ \text{min}^{-1}$, $\sigma_u^- = 0.003\ \text{min}^{-1}$, $\sigma_u^+ = 0.71\sigma_u^-$, $\lambda = 2d$, and $\alpha = 0$. See *SI Appendix*, section 9 for details.

replication, binomial partitioning due to cell division, dosage compensation, growth-dependent transcription, promoter switching, and the translation of mature mRNA into proteins. The model also provides a description of both nascent and mature mRNA distributions, which is necessary to make sense of high-resolution experimental data (11). We have shown that by breaking this complex model into a set of simpler submodels, solving each submodel approximately using timescale-separation methods (50), and then integrating the results together, it is possible to derive closed-form expressions for the time-dependent distribution of the numbers of nascent mRNA, mature mRNA, and protein inside a single cell. Specifically, we have made use of two assumptions: 1) Nascent mRNA dynamics is much faster than mRNA dynamics, which itself is much faster than protein dynamics; and 2) the gene-inactivation rate is much larger than the gene-activation rate. We have provided experimental evidence that these assumptions are reasonable for a large number of genes in several different types of eukaryotic cells grown under different conditions, and, hence, our model provides a detailed quantitative model of eukaryotic gene expression. A major advantage of our analytic approach is that, despite the biological complexity described by the model, it leads to simple distributions (negative binomial) for the molecule numbers for all cell ages and generations. It also provides a quantitative description for both nascent and mature mRNA dynamics, both of which are measurable observables. A description in terms of the two is advantageous, since nascent mRNA closely reflects the kinetics of transcription, while mature mRNA reflects additional processes downstream of transcription.

Numerical evaluation of these distributions is far more computationally efficient than direct simulation using the SSA. This implies that the model's behavior can be easily predicted across vast swaths of parameter space and that usually prohibitive tasks, such as stochastic sensitivity analysis, can be straightforwardly performed. Indeed, using our theory, we calculated the sensitivity of the coefficient of variation of noise (averaged over the cell cycle and in the cyclo-stationary limit) to small changes in the parameter values measured for mammalian cells. The parameters ordered according to the magnitude of their sensitivities are mRNA degradation rate, the rate of gene activation (before and after replication), rate of gene inactivation, transcription rate, cell-division time, replication time, and the parameter determining the coupling between transcription rate and cell growth. This suggests that variations in the values of the mRNA degradation rates and of the promoter-switching rates across cells are among the most significant sources of variability in gene expression across a population of cells (what is often termed extrinsic noise). Another major advantage of our closed-form expressions for the time-dependent distributions, and, in particular, their approximation by negative binomial distributions, is that they can be used to obtain the likelihood of a set of experimental observations of the molecule numbers—the likelihood can then be used within a Markov chain Monte Carlo algorithm to obtain the posterior distributions of parameters (51, 52).

Our model cannot resolve the effects of polymerase and transcription-factor fluctuations on mRNA and protein dynamics (53). It is also the case that our model cannot take into account the effect of cell-cycle-length variability on the distribution of protein numbers because the low protein degradation rates do not average out timing fluctuations (54). Lastly, the sharing of resources can potentially modify many cellular processes (55). Future work will involve extensions of the model to include these and other salient features of single-cell biology.

Data Availability. The data used in the paper are described in *SI Appendix*, section 2. The simulation code is available from the corresponding author upon request.

ACKNOWLEDGMENTS. Z.C. was supported by the UK Research Councils' Synthetic Biology for Growth program, the Biotechnology and Biological Sciences Research Council (BBSRC), the Engineering and Physical Sciences Research Council, and Medical Research Council Grant BB/M018040/1. R.G. was supported by BBSRC Grant BB/M025551/1. R.G. thanks Sara Buonomo and Peter Swain for useful discussions and insightful feedback.

1. M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
2. J. Paulsson, Summing up the noise in gene networks. *Nature* **427**, 415–418 (2004).
3. V. Shahrezaei, P. S. Swain, Analytical distributions for stochastic gene expression. *Proc. Acad. Natl. Sci. U.S.A.* **105**, 17256–17261 (2008).
4. D. Schnoerr, G. Sanguinetti, R. Grima, Approximation and inference methods for stochastic biochemical kinetics: A tutorial review. *J. Phys. A* **50**, 093001 (2017).
5. N. Battich, T. Stoeger, L. Pelkmans, Control of transcript variability in single mammalian cells. *Cell* **163**, 1596–1610 (2015).
6. N. Maheshri, E. K. O'Shea, Living with noisy genes: How cells function reliably with inherent variability in gene expression. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 413–434 (2007).
7. I. Golding, J. Paulsson, S. M. Zawilski, E. C. Cox, Real-time kinetics of gene activity in individual bacteria. *Cell* **123**, 1025–1036 (2005).
8. A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, S. Tyagi, Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).
9. R. D. Dar *et al.*, Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Acad. Natl. Sci. U.S.A.* **109**, 17454–17459 (2012).
10. A. J. M. Larsson *et al.*, Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254 (2019).
11. D. Zenklusen, D. R. Larson, R. H. Singer, Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat. Struct. Mol. Biol.* **15**, 1263–1271 (2008).
12. G. La Manno *et al.*, RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
13. J. Peccoud, B. Ycart, Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).
14. D. M. Suter *et al.*, Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472–474 (2011).
15. N. Molina *et al.*, Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. *Proc. Acad. Natl. Sci. U.S.A.* **110**, 20563–20568 (2013).
16. L.-h. So *et al.*, General properties of transcriptional time series in *Escherichia coli*. *Nat. Genet.* **43**, 554–560 (2011).
17. M. Kaern, T. C. Elston, W. J. Blake, J. J. Collins, Stochasticity in gene expression: From theories to phenotypes. *Nat. Rev. Genet.* **6**, 451–464 (2005).
18. A. Senecal *et al.*, Transcription factors modulate c-Fos transcriptional bursts. *Cell Rep.* **8**, 75–83 (2014).
19. S. O. Skinner *et al.*, Single-cell analysis of transcription kinetics across the cell cycle. *Elife* **5**, e12175 (2016).
20. O. G. Berg, A model for the statistical fluctuations of protein numbers in a microbial population. *J. Theor. Biol.* **71**, 587–603 (1978).
21. D. Huh, J. Paulsson, Random partitioning of molecules at cell division. *Proc. Acad. Natl. Sci. U.S.A.* **108**, 15004–15009 (2011).
22. J. R. Peterson, J. A. Cole, J. Fei, T. Ha, Z. A. Luthey-Schulten, Effects of DNA replication on mRNA noise. *Proc. Acad. Natl. Sci. U.S.A.* **112**, 15886–15891 (2015).
23. O. Padovan-Merhar *et al.*, Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol. Cell* **58**, 339–352 (2015).
24. J. Lin, A. Amir, Homeostasis of protein and mRNA concentrations in growing cells. *Nat. Commun.* **9**, 4496 (2018).
25. D. T. Gillespie, Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).
26. H. Ochiai, T. Sugawara, T. Sakuma, T. Yamamoto, Stochastic promoter activation affects NANOG expression variability in mouse embryonic stem cells. *Sci. Rep.* **4**, 7125 (2014).
27. K. B. Halpern *et al.*, Bursty gene expression in the intact mammalian liver. *Mol. Cell* **58**, 147–156 (2015).
28. H. Xu, S. O. Skinner, A. M. Sokac, I. Golding, Stochastic kinetics of nascent RNA. *Phys. Rev. Lett.* **117**, 128101 (2016).
29. Y. Wang *et al.*, Precision and functional specificity in mRNA decay. *Proc. Acad. Natl. Sci. U.S.A.* **99**, 5860–5865 (2002).
30. C. Cadart *et al.*, Size control in mammalian cells involves modulation of both growth rate and cell cycle duration. *Nat. Commun.* **9**, 3275 (2018).
31. N. Reverón-Gómez *et al.*, Accurate recycling of parental histones reproduces the histone modification landscape during DNA replication. *Mol. Cell* **72**, 239–249 (2018).
32. S. Hamperl, K. A. Cimprich, Conflict resolution in the genome: How transcription and replication make it work. *Cell* **167**, 1455–1467 (2016).
33. C. A. Vargas-Garcia, K. R. Ghusinga, A. Singh, Cell size control and gene expression homeostasis in single-cells. *Curr. Opin. Struct. Biol.* **8**, 109–116 (2018).

34. O. Sandler *et al.*, Lineage correlations of single cell division time as a probe of cell-cycle dynamics. *Nature* **519**, 468–471 (2015).
35. A. Golubev, Applications and implications of the exponentially modified gamma distribution as a model for time variabilities related to cell proliferation and gene expression. *J. Theor. Biol.* **393**, 203–217 (2016).
36. G. M. Walker, Synchronization of yeast cell populations. *Methods Cell Sci.* **21**, 87–93 (1999).
37. Y. Tian, C. Luo, Y. Lu, C. Tang, Q. Ouyang, Cell cycle synchronization by nutrient modulation. *Integr. Biol.* **4**, 328–334 (2012).
38. C. Gérard, A. Goldbeter, Entrainment of the mammalian cell cycle by the circadian clock: Modeling two coupled cellular rhythms. *PLoS Comput. Biol.* **8**, e1002516 (2012).
39. E. Kowalska *et al.*, NONO couples the circadian clock to the cell cycle. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 1592–1599 (2013).
40. P. H. O'Farrell, J. Stumpff, T. T. Su, Embryonic cleavage cycles: How is a mouse like a fly?*Curr. Biol.* **14**, R35–R45 (2004).
41. P. Thomas, Making sense of snapshot data: Ergodic principle for clonal cell populations. *J. R. Soc. Interface* **14**, 20170467 (2017).
42. C. A. Yates, M. J. Ford, R. L. Mort, A multi-stage representation of cell proliferation as a Markov process. *Bull. Math. Biol.* **79**, 2905–2928 (2017).
43. B. Ingalls, Sensitivity analysis: From model parameters to system behaviour. *Essays Biochem.* **45**, 177–194 (2008).
44. A. A. Cohen *et al.*, Protein dynamics in individual human cells: Experiment and theory. *PLoS One* **4**, e4901 (2009).
45. A. J. Hughes *et al.*, Single-cell Western blotting. *Nat. Methods* **11**, 749–755 (2014).
46. C. Albayrak *et al.*, Digital quantification of proteins and mRNA in single mammalian cells. *Mol. Cell* **61**, 914–924 (2016).
47. B. Schwanhäusser *et al.*, Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
48. R. Grima, D. R. Schmidt, T. J. Newman, Steady-state fluctuations of a genetic feedback loop: An exact solution. *J. Chem. Phys.* **137**, 035104 (2012).
49. Z. Cao, R. Grima, Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nat. Commun.* **9**, 3305 (2018).
50. D. Cappelletti, C. Wiuf, Elimination of intermediate species in multiscale stochastic reaction networks. *Ann. Appl. Probab.* **26**, 2915–2958 (2016).
51. V. Stathopoulos, M. A. Girolami, Markov chain Monte Carlo inference for Markov jump processes via the linear noise approximation. *Philos. Trans. Roy. Soc. A* **371**, 20110541 (2013).
52. Z. Cao, R. Grima, Accuracy of parameter estimation for auto-regulatory transcriptional feedback loops from noisy data. *J. R. Soc. Interface* **16**, 20180967 (2019).
53. C. R. Bartman *et al.*, Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation. *Mol. Cell* **73**, 519–532 (2019).
54. M. Soltani, C. A. Vargas-Garcia, D. Antunes, A. Singh, Intercellular variability in protein levels from stochastic expression and noisy cell cycle processes. *PLoS Comput. Biol.* **12**, e1004972 (2016).
55. A. Y. Weiße, D. A. Oyarzún, V. Danos, P. S. Swain, Mechanistic links between cellular trade-offs, gene expression, and growth. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E1038–E1047 (2015).